DATA400 – Idea 2 – Kevin Tran

1. General ideas

- *The* **Question**: Can the discussions on social media platforms predict short-term price volatility for specific, highly-discussed "meme stocks"?
- This idea originates from the fact that in 2024, Elon Musk's DOGE coin price fluctuates based on his Twitter (X) posts.
- My sequence is in ECON so I think this might be relatable.

2. Detailed

- This project will combine Natural Language Processing (NLP) with time-series analysis.
    - Social Media Data (Independent Variable):
        - Data will be scraped from relevant online forums, particularly Reddit communities.
        - Using Python libraries, posts and comments related to the target stocks will be collected.
    - Financial Data (Dependent Variable):
        - Historical daily stock data (opening price, closing price, volume) will be sourced from a public API, such as Yahoo Finance.
    - Analytical Model:
        - Sentiment Analysis: An NLP model will be used to assign a sentiment score to each post and comment.
        - Time-Series Creation: The sentiment scores will be aggregated daily to create a "Daily Sentiment Score" for each stock.
        - Predictive Modeling: A time-series analysis (such as a VAR model) will be performed to determine if the sentiment score time series has any statistically significant power in predicting the next day's stock price volatility or trading volume.
- This project is highly relevant in the modern economy and offers several benefits:
    - Practical ML Application: It provides a opportunity to implement a full data science pipeline, from web scraping and data cleaning to advanced machine learning and statistical analysis.
    - Behavioral Economics: The study directly relates to your economics sequence by exploring how herd behavior and public sentiment-rather than just company fundamentals-can act as powerful drivers in financial markets.
    - Innovative Research: This topic is at the forefront of financial analysis, offering a chance to contribute to the body of knowledge on how social media is transforming markets.
- The stakes are high when dealing with financial markets, and the ethical lines must be clearly drawn.
    - Stakes: The primary risk is the misinterpretation of the model as a financial tool. A flawed or over-fitted model could falsely suggest a reliable way to

"beat the market." If the results were miscommunicated, they could implicitly encourage uncontrolled financial behavior. This project is an academic analysis, not the creation of a trading algorithm.

- Ethical Implications:
  - No Financial Advice: The final report must contain a strong disclaimer stating that the findings are for academic purposes only, are based on historical data, and do not constitute financial advice.
  - Data Anonymization: When scraping data, all usernames and personally identifiable information must be removed to protect the privacy of the forum users.
  - Acknowledging Manipulation: The project must acknowledge that the platforms being studied can be subject to intentional manipulation (e.g., "pump and dump" schemes). The model is analyzing sentiment, but it cannot determine the intent behind that sentiment.